A simulation study of maximum likelihood estimation in logistic regression with cured individuals

Aba Diop, Aliou Diop, Jean-François Dupuy

A simulation study of maximum likelihood estimation in logistic regression with cured individuals

Aba DIOP

MIA, Université de La Rochelle, France and LERSTAD, Université Gaston Berger, Saint Louis, Sénégal.
Email: aba.diop@univ-lr.fr

Aliou DIOP

LERSTAD, Université Gaston Berger, Saint Louis, Sénégal.
Email: aliou.diop@ugb.edu.sn

Jean-François DUPUY†

IRMAR-Institut National des Sciences Appliquées de Rennes, France.
Email: Jean-Francois.Dupuy@insa-rennes.fr

Abstract. The logistic regression model is widely used to investigate the relationship between a binary outcome Y and a set of potential predictors $\mathbf{X}$. Diop *et al.* (2011) present some conditions under which the maximum likelihood estimator is consistent and asymptotically normal in the logistic regression model with a cure fraction. So far, however, only limited simulation results are available to judge the quality of this estimator in finite samples. Therefore in this paper, we conduct a detailed simulation study of its numerical properties. We evaluate its accuracy and the quality of the normal approximation of its asymptotic distribution. We also study the quality of the approximation for constructing asymptotic Wald-type tests of hypothesis. Finally, we consider the problem of estimating the conditional probability of the outcome. Our results indicate that when the proportion of cured individuals is moderate to moderately large, and the sample size is large enough, reliable statistical inferences can be obtained for the regression effects and the probability of the outcome. Our results also indicate that the approximations can be problematic when the cure fraction is very large.

Keywords: Zero-inflation, mixture model, maximum likelihood estimation

1. Introduction

Logistic regression has become a standard tool to investigate the relationship between a binary response Y and a set of potential predictors $\mathbf{X}$. In the medical setting for example, the response may represent the infection status with respect to some disease. If Y_i denotes the infection status for the i-th individual in a sample of size n ($Y_i = 1$ if the individual is infected, and $Y_i = 0$ otherwise) and $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})'$ is the corresponding predictor, logistic regression models the conditional probability $\mathbb{P}(Y_i = 1|\mathbf{X}_i)$ of infection as

$$\log\left(\frac{\mathbb{P}(Y_i = 1|\mathbf{X}_i)}{1 - \mathbb{P}(Y_i = 1|\mathbf{X}_i)}\right) = \beta'\mathbf{X}_i, \tag{1}$$

†Corresponding author

where $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown regression parameter to be estimated. Estimation and testing procedures in the model (1) are well established (see, for example, Hosmer and Lemeshow (2000) and Hilbe (2009)) and are available in standard statistical softwares. In particular, the maximum likelihood estimator (MLE) of β was shown to be consistent, and approximately normally distributed in large samples (Gouriéroux and Monfort (1981), Fahrmeir and Kaufmann (1985)).

Recently, Diop *et al.* (2011) have considered the problem of estimation in the model (1) when there is a cured fraction in the sample. In medical studies, it often arises that a proportion of the study subjects cannot experience the outcome of interest. Such individuals are said to be cured, or immune. The population under study can then be considered as a mixture of cured and susceptible subjects, where a subject is said to be susceptible if he would eventually experience the outcome of interest. One problem arising in this setting is that it is usually unknown who are the susceptible, and the cured subjects (unless the outcome of interest has been observed). Consider, for example, the occurrence of infection from some disease to be the outcome of interest. Then, if a subject is uninfected, the investigator usually does not know whether this subject is immune to the infection, or susceptible albeit still uninfected.

Estimating a regression model with a cure fraction can be viewed as a zero-inflated regression problem. Zero-inflation occurs in the analysis of count data when the observations contain more zeros than expected. Failure to account for these extra zeros is known to result in biased parameter estimates and inferences. Motivated by various applications in public health, epidemiology, sociology, engineering, agriculture, a variety of zero-inflated regression models have been developed and extended, such as the zero-inflated Poisson (ZIP) model (see, among others, Lambert (1992), Dietz and Böhning (2000), Ridout *et al.* (2001), Lam *et al.* (2006), Xiang *et al.* (2007)), the zero-inflated binomial (ZIB) model (see Hall (2000)), the zero-inflated negative binomial (ZINB) model (see Moghimbeigi *et al.* (2008)). Various other models and numerous references can be found in Famoye and Singh (2006), Lee *et al.* (2006), Kelley and Anderson (2008), and Moghimbeigi *et al.* (2009).

By assuming a logistic regression model for the probability of being cured, Diop *et al.* (2011) developed a new zero-inflated Bernoulli regression model with logit links for both the binary response of interest (the probability of infection, say), and the zero-inflation probability (of being cured). The authors investigated, mainly theoretically, the issues of identifiability and estimation in this model. In particular, they proposed a set of sufficient conditions for model identifiability, and they proved the consistency and asymptotic normality of the MLE.

In this paper, as a supplement of the theoretical work of Diop *et al.* (2011), we conduct a detailed simulation study of the numerical behavior of their estimator. Precisely, we evaluate the influence of various factors (sample size, proportion of immunes in the sample, proportion of infected among the susceptibles) on the accuracy of the estimator and on the quality of the Gaussian approximation of its asymptotic distribution. The performance (in terms of power and level) of the Wald-type test of "$\beta = 0$" is also investigated. We also consider the problem of estimating the probability of infection $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$, for some value $\mathbf{x}$ of the covariates.

The rest of the paper is organized as follows. In the Section 2, we describe the model and the estimation procedure proposed by Diop *et al.* (2011), and we recall the theoretical properties of the resulting estimator of the regression parameter β. We also construct asymptotic confidence intervals for a probability of infection $p(\mathbf{x})$. Section 3 describes the detailed simulation study. A discussion and some conclusions and perspectives are given in

the Section 4.

2. Model and estimation

Let $\mathcal{O}_i = (Y_i, S_i, \mathbf{X}_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ be independent copies of the random vector $\mathcal{O} = (Y, S, \mathbf{X}, \mathbf{Z})$, where for every i, Y_i is a binary response indicating the infection status (with respect to some disease) of the i-th individual (that is, $Y_i = 1$ if the individual i is infected, and $Y_i = 0$ otherwise), and S_i is a binary variable indicating whether the individual is susceptible to the infection ($S_i = 1$) or immune ($S_i = 0$). If $Y_i = 0$, then the value of S_i is unknown. $\mathbf{X}_i = (1, X_{i2}, \ldots, X_{ip})'$ and $\mathbf{Z}_i = (1, Z_{i2}, \ldots, Z_{iq})'$ are covariate vectors related to the infection status and immunity respectively. $\mathbf{X}_i$ and $\mathbf{Z}_i$ may contain quantitative and qualitative components, and may even share some components.
The zero-inflated Bernoulli regression model described by Diop *et al.* (2011) is defined by the following equations for the infection status:

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta'\mathbf{X}_i & \text{if } \{S_i = 1\} \\ \mathbb{P}(Y = 1|\mathbf{X}_i, S_i) = 0 & \text{if } \{S_i = 0\} \end{cases} \tag{2}$$

and by the following model for immunity:

$$\log\left(\frac{\mathbb{P}(S = 1|\mathbf{Z}_i)}{1 - \mathbb{P}(S = 1|\mathbf{Z}_i)}\right) = \theta'\mathbf{Z}_i. \tag{3}$$

In this model, $\beta = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ is an unknown regression parameter of interest (β measures the association between the potential predictors $\mathbf{X}_i$ and the risk of infection for a susceptible individual), and $\theta = (\theta_1, \ldots, \theta_q)' \in \mathbb{R}^q$ is an unknown nuisance parameter. Letting $\psi := (\beta', \theta')'$, the log-likelihood for ψ from the sample $\mathcal{O}_1, \ldots, \mathcal{O}_n$ (where S_i is unknown when $Y_i = 0$) is

$$\begin{aligned} l_n(\psi) &= \sum_{i=1}^{n} \Big\{ Y_i(\beta'\mathbf{X}_i + \theta'\mathbf{Z}_i) + (1 - Y_i)\log(1 + e^{\beta'\mathbf{X}_i} + e^{\theta'\mathbf{Z}_i}) \\ &\qquad - \log\left(1 + e^{\beta'\mathbf{X}_i}\right) - \log\left(1 + e^{\theta'\mathbf{Z}_i}\right) \Big\} \\ &:= \sum_{i=1}^{n} l(\psi; \mathcal{O}_i). \end{aligned}$$

The MLE $\widehat{\psi}_n := (\widehat{\beta}_n', \widehat{\theta}_n')'$ of ψ is defined as the solution of the score equation $\partial l_n(\psi)/\partial\psi = 0$, which can be solved, for example, using the `optim` function of the software `R`.
The main assumptions underlying the asymptotic results proved by Diop *et al.* (2011) are the following:

A1 The covariates are bounded. For every $i = 1, 2, \ldots$, $j = 2, \ldots, p$, $k = 2, \ldots, q$, $\text{var}[X_{ij}] > 0$ and $\text{var}[Z_{ik}] > 0$. For every $i = 1, 2, \ldots$, the X_{ij} ($j = 1, \ldots, p$) are linearly independent, and the Z_{ik} ($k = 1, \ldots, q$) are linearly independent.

A2 There exists a continuous covariate V which is in $\mathbf{X}$ but not in $\mathbf{Z}$ that is, if β_V and θ_V denote the coefficients of V in the linear predictors (2) and (3) respectively, then $\beta_V \neq 0$ and $\theta_V = 0$. At a model-building stage, it is known that V is in $\mathbf{X}$.

Remark 1. The condition A1 is a classical condition for identifiability and asymptotic results in standard logistic regression (see, for example, Gouriéroux and Monfort (1981)). The condition A2, which imposes some restrictions on the covariates, is required for the identifiability of ψ in the joint model (2)-(3) (we may alternatively assume that the continuous covariate V is in $\mathbf{Z}$ but not in $\mathbf{X}$). In the following, we will assume that V is in $\mathbf{X}$ but not in $\mathbf{Z}$, with $\beta_V := \beta_l$ for some $l \in \{2, \ldots, p\}$, and for the i-th individual, we will denote V_i by X_{il}. We refer to Diop *et al.* (2011) for a detailed discussion about this condition.

Let $\mathcal{I}_\psi = -\mathbb{E}[\partial^2 l(\psi; \mathcal{O})/\partial\psi\partial\psi']$. Under the conditions stated above, Diop *et al.* (2011) prove the following result:

Theorem. *As n tends to infinity, $\sqrt{n}(\widehat{\psi}_n - \psi)$ converges in distribution to a zero-mean Gaussian vector, whose covariance matrix $\mathcal{I}_\psi^{-1}$ can be consistently estimated by $\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}$, where $\widehat{\mathcal{I}}_{\widehat{\psi}_n} = -n^{-1}(\partial^2 l_n(\psi)/\partial\psi\partial\psi')|_{\psi=\widehat{\psi}_n}$.*

The parameter of interest in the model (2)-(3) is β. The asymptotic distribution of $\widehat{\beta}_n$ now easily follows:

Corollary 1. *Let M be the $(p \times (p+q))$ block-matrix $[I_p, 0_{p,q}]$, where I_p denotes the identity matrix of order p and $0_{p,q}$ is the $(p \times q)$ matrix whose components are all equal to 0. Then $\sqrt{n}(\widehat{\beta}_n - \beta)$ converges in distribution to a zero-mean Gaussian vector with covariance matrix $M\mathcal{I}_\psi^{-1}M'$, which is the upper-left $(p \times p)$ block of $\mathcal{I}_\psi^{-1}$.*

The convergence in distribution of $\widehat{\beta}_n$ can now be used to make statistical inference about β. For example, if one wishes to test the null hypothesis $H_0 : \beta_l = 0$ against the alternative $H_1 : \beta_l \neq 0$ (for some $1 \leq l \leq p$), one can rely on the usual Wald-type test, which rejects H_0 at the asymptotic level α $(0 < \alpha < 1)$ if

$$\left| \frac{\widehat{\beta}_{n,l}}{\sqrt{\frac{(M\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}M')_{ll}}{n}}} \right| > u_{1-\frac{\alpha}{2}},$$

where $u_{1-\frac{\alpha}{2}}$ is the quantile of order $(1-\frac{\alpha}{2})$ of the standard normal law, $\widehat{\beta}_{n,l}$ is the l-th component of $\widehat{\beta}_n$, and $(M\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}M')_{ll}$ denotes the l-th diagonal component of $M\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}M'$. This corollary can also be used to estimate a given probability $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$ of infection (this result was not mentioned by Diop *et al.* (2011)):

Corollary 2. *Let $\mathbf{x}$ be a given value of the covariate $\mathbf{X}$. As n tends to infinity, $\sqrt{n}(\widehat{\beta}_n - \beta)'\mathbf{x}$ converges in distribution to a zero-mean Gaussian random variable with variance $\mathbf{x}'M\mathcal{I}_\psi^{-1}M'\mathbf{x}$. An asymptotic $(1-\alpha)$-level confidence interval for $p(\mathbf{x})$ is therefore*

$$\left[\frac{e^{u_n(\mathbf{x})}}{1+e^{u_n(\mathbf{x})}}, \frac{e^{v_n(\mathbf{x})}}{1+e^{v_n(\mathbf{x})}} \right],$$

where

$$u_n(\mathbf{x}) = \widehat{\beta}_n'\mathbf{x} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbf{x}'M\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}M'\mathbf{x}}{n}} \text{ and } v_n(\mathbf{x}) = \widehat{\beta}_n'\mathbf{x} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\mathbf{x}'M\widehat{\mathcal{I}}_{\widehat{\psi}_n}^{-1}M'\mathbf{x}}{n}}.$$

In a limited simulation study, Diop *et al.* (2011) have investigated the numerical behavior of $\widehat{\beta}_n$ for some very simple models. In the present paper, we undertake a much more detailed numerical analysis of this estimator, focusing on various measures of its accuracy (namely the bias, and the root mean-square and mean absolute errors), on the quality of the Gaussian approximation of its asymptotic distribution, and on the accuracy of an estimator of the probability of infection $p(\mathbf{x})$, for some $\mathbf{x}$.

3. Simulation study

3.1. Study design

The simulation setting is as follows. We consider the following models for the infection status:

$$\begin{cases} \log\left(\frac{\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}{1-\mathbb{P}(Y=1|\mathbf{X}_i,S_i)}\right) = \beta_1 + \beta_2 X_{i2} + \beta_3 Z_{i2} + \beta_4 Z_{i3} + \beta_5 Z_{i4} & \text{if } S_i = 1 \\ \mathbb{P}(Y=1|\mathbf{X}_i,S_i) = 0 & \text{if } S_i = 0 \end{cases} \tag{4}$$

and the immunity status:

$$\log\left(\frac{\mathbb{P}(S=1|\mathbf{Z}_i)}{1-\mathbb{P}(S=1|\mathbf{Z}_i)}\right) = \theta_1 + \theta_2 Z_{i2} + \theta_3 Z_{i3} + \theta_4 Z_{i4}, \tag{5}$$

where $X_{i2} \sim N(0,1)$, $Z_{i2} \sim N(1,1)$, and Z_{i3} and Z_{i4} are indicator variables built from a categorical variable with 3 categories. Note that X_2 plays the role of the continuous covariate V in the condition A2. Note also that the models for infection and immunity share no less than three covariates, including both continuous and discrete variables (this is the least favorable case with respect to the identifiability of the parameters). An i.i.d. sample of size n of the vector $(Y, S, \mathbf{X}, \mathbf{Z})$ is generated from this model, and for each individual i, we get a realization $(y_i, s_i, \mathbf{x}_i, \mathbf{z}_i)$, where s_i is considered as unknown if $y_i = 0$. The maximum likelihood estimator $\widehat{\beta}_n$ of $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ is obtained from this incomplete dataset by solving the score equation given in Section 2, using the `optim` function in `R`. As a by-product of the method, an estimate is also obtained for the nuisance parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)'$.

The finite-sample behavior of the maximum likelihood estimator $\widehat{\beta}_n$ is assessed for several sample sizes ($n = 100, 500, 1000, 1500$) and various values for the percentage of immunes in the sample, namely 25%, 50%, and 75%. We also consider different values for the proportion of infected individuals among the susceptibles (30% and 70%). The desired proportions of immunes and infected are obtained by choosing appropriate values for β and θ. The following values are considered for β:

- model $\mathcal{M}_1$: $\beta = (-1.7, -2, -3.4, 5, 0.3)'$. Using these values, approximately 30% of the susceptibles are infected.
- model $\mathcal{M}_2$: $\beta = (1.5, -2.3, 2.5, -3.5, 0.5)'$. Approximately 70% of the susceptibles are infected.
- model $\mathcal{M}_3$: $\beta = (-1.7, -2.8, 0, -0.7, 1.1)'$. Approximately 30% of the susceptibles are infected.

- model $\mathcal{M}_4$: $\beta = (1.5, -2, 0, 3.5, -4)'$. Approximately 70% of the susceptibles are infected.

Letting $\beta_3 = 0$ (models $\mathcal{M}_3$ and $\mathcal{M}_4$) implies that Z_2 does not influence the risk of infection. However, when estimating the model (4)-(5) using Diop *et al.* (2011)'s procedure, we use the whole set of covariates (X_2, Z_2, Z_3, Z_4). This allows us to evaluate the level of a Wald-type test of nullity of β_3.

3.2. Results

For each configuration `sample size` $\times$ `percentage of immunes` $\times$ `percentage of infected among susceptibles` of the design parameters, $N = 1500$ samples are obtained. Based on these N repetitions, we obtain averaged values for the estimates of the β_l $(l = 1, \ldots, 5)$, which are calculated as $N^{-1}\sum_{j=1}^{N}\widehat{\beta}_{l,n}^{(j)}$, where $\widehat{\beta}_n^{(j)} = (\widehat{\beta}_{1,n}^{(j)}, \ldots, \widehat{\beta}_{5,n}^{(j)})'$ is the estimate obtained from the j-th simulated sample. For each of the parameters β_l, we also obtain the empirical root mean square (RMSE) and mean absolute errors (MAE), based on the N samples. Similar results are obtained for the nuisance parameter θ. When $\beta_l \neq 0$ (respectively $\beta_l = 0$), we obtain the empirical power (respectively the empirical size) of the Wald test at the 5% level for testing $H_0 : \beta_l = 0$. The null hypothesis $H_0 : \beta_l = 0$ is the hypothesis that the predictor X_l does not influence the risk of infection of susceptible individuals.
Tables 1 and 2 give the results for the models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. There, $(\cdot)$ indicates the root mean square error, $[\cdot]$ indicates the mean absolute error, and * is the empirical power of the Wald test.

Table 1 about here

Table 2 about here

In order to investigate the level of Wald-type tests of hypothesis of the form $H_0 : \beta_l = 0$, we simulate samples from the models $\mathcal{M}_3$ and $\mathcal{M}_4$. The results are given in the tables 3 and 4. In these tables, $*$ indicates the empirical power of the Wald test (when $\beta_l \neq 0$) and $\dagger$ indicates the empirical size (when $\beta_l = 0$).

Table 3 about here

Table 4 about here

If there were no immunes in the sample, a usual logistic regression model could be fitted to the data. The results for such an hypothetical case are interesting since they provide a benchmark for evaluating the performance of the proposed estimator. Tables 5 and 6 provide the results for the models $\mathcal{M}_1$ and $\mathcal{M}_3$ respectively, if there were no immunes in the samples. The results for the other models yield similar conclusions and are therefore omitted.

Table 5 about here

Table 6 about here

From all these results, it appears that the maximum likelihood estimator proposed by Diop *et al.* (2011) provides a reasonable approximation of the true regression parameter, even

when the immune fraction is high. Overall, the bias of $\widehat{\beta}_n$ stays limited while its variability increases with the percentage of immunes. This increase is particularly noticeable when the sample size is small ($n = 100$) and when the sample size is moderate ($n = 500$) with a high percentage of immunes (75%). As a consequence, the power of Wald tests of nullity of regression coefficients can be low (compared to the case where there are no immunes) when the percentage of immunes is very high and/or the sample size is small. For moderately large to large sample sizes ($n \geq 500$, say), the power of the Wald test indicates good numerical performance of the inferential procedure proposed by Diop *et al.* (2011), provided that the magnitude of the regression coefficient is sufficiently large. The level of the Wald test is globally respected when the percentage of immunes is moderate (25%) and degradates when this percentage increases.

Then, we compare these results to the ones obtained from a "naive" method where: i) we consider every individual i such that $\{Y_i = 0\}$ as being susceptible but uninfected (that is, we ignore the eventual immunity of this individual) and ii) we apply a usual logistic regression analysis to the resulting dataset. The results of such "naive" analysis for the model $\mathcal{M}_1$ are given in the table 7. The results for the other models yield similar observations and thus, they are not given here.

Table 7 about here

From this table, it appears that missing the immunity present in the sample results in strongly biased estimates of the true association between the covariates and the risk of infection. The variability of the estimates is also very important, even for very large sample sizes.

In logistic regression, it is usually of interest to estimate the probability of infection $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$, for some given value $\mathbf{x}$ of the covariates. An obvious estimate of $p(\mathbf{x})$ is $\widehat{p}_n(\mathbf{x}) := \exp(\widehat{\beta}_n'\mathbf{x})/(1 + \exp(\widehat{\beta}_n'\mathbf{x}))$. In table 8, we investigate the numerical properties of this estimator (restraining ourselves to one value of $p(\mathbf{x})$ for each of the models $\mathcal{M}_1$ and $\mathcal{M}_2$). For every combination of the simulation design parameters, we obtain the average estimated probability and the corresponding RMSE and MAE. We also provide the empirical coverage probability of asymptotic 95%-confidence intervals for $p(\mathbf{x})$, and the average length of these intervals.

Table 8 about here

From this table, it appears that $\widehat{p}_n(\mathbf{x})$ provides a reasonable estimate of $p(\mathbf{x})$ even when the immune fraction is high. The coverage probabilities and average interval lengths indicate good performance (compared to the no-immunes case) of the confidence intervals for $p(\mathbf{x})$ obtained from Diop *et al.* (2011)'s results, when the cure fraction is moderate (even for $n = 100$), and when the cure fraction is moderately large (50%) and the sample size is large ($n \geq 1000$).

We also investigate the quality of the normal approximation of the large-sample distribution of $\widehat{\beta}_n$. For each configuration of the design parameters, we obtain histograms of the $\widehat{\beta}_{n,l}^{(j)}$ ($j = 1, \ldots, N$), along with the corresponding Q-Q plots. The results are provided for the coefficients β_2 and β_3 of the model $\mathcal{M}_1$. The results for the other parameters and models yield similar observations and are therefore omitted.

Figures 1 to 8 about here

From these figures, the normal approximation seems to be reasonably satisfied when the immune fraction is moderate (25%) and the sample size is sufficiently large ($n \geq 500$). When the proportion of immunes attains 50%, a larger sample size ($n \geq 1000$, say) is required to consider this approximation as being still reasonable. When the immune proportion is very large (75%), the distribution of the estimator can be highly skewed, particularly when the sample size is small.

Finally, we investigate the quality of the normal approximation of the distribution of $\widehat{p}_n(\mathbf{x})$. Histograms and Q-Q plots of the $\widehat{p}_n^{(j)}(\mathbf{x})$ ($j = 1, \ldots, N$) are obtained for one value of $p(\mathbf{x})$ in the model $\mathcal{M}_1$, and are given in Figures 9-12.

Figures 9 to 12 about here

From these figures, the normal approximation is reasonably satisfied when the immune fraction is moderate to moderately large (25%-50%) provided that the sample size is sufficiently large ($n \geq 1000$, say). The distribution of $\widehat{p}_n(\mathbf{x})$ can be highly skewed otherwise.

Overall, these results indicate that a reliable statistical inference on the regression effects and probabilities of event in the logistic regression model with a cure fraction should be based on a sample having, at least, a moderately large size ($n \geq 500$, say) when the immune fraction is low (25%), or a large size ($n \geq 1000$) when the immunity attains 50% of the sample. When the immune proportion is very large (75%), the results should be considered carefully, considering the increase in the variability of the estimates and the skewness of their distributions.

4. Discussion

In this paper, we have conducted a detailed numerical investigation of the maximum likelihood estimators proposed by Diop *et al.* (2011) for estimating the logistic regression model when there is a cure fraction in the sample. This problem can also be viewed as a problem of zero-inflation in the logistic regression model, or equivalently, as a problem of statistical inference in logistic regression from a mixture of cured and susceptible individuals. From our investigations, the maximum likelihood estimators of both the regression effects of interest and probabilities of event (such as infection, say) perform quite well under appropriate conditions regarding the sample size and immune proportion. Overall, reliable statistical inferences (point estimation, confidence intervals, hypothesis testing) should be obtained when the immune fraction is moderate (25%) and the sample size is at least 500, or the immune fraction is moderately large (50%) and the sample size is at least 1000.

Several open problems still deserve attention. First, our results were obtained under the assumption that the cure model is correctly specified. It is now of interest to investigate the effect on the whole inference of a misspecification of this model. One may also consider estimating the cure model by using more flexible methods, such as nonparametric kernel regression. Another stimulating topic of interest is to consider the problem of immunes in the more general semiparametric logistic regression model. Both topics are the subject for our future research.

In logistic regression, it is sometimes of interest to make inference about the probability of event $\mathbb{P}(Y = 1|\mathbf{X}, S = 1)$ across the whole range of the predictors $\mathbf{X}$. Pointwise confidence intervals are not adequate for that purpose. The calculation of simultaneous confidence bands for the probabilities $\{p(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ thus constitutes another issue of (both methodological and practical) interest. Several methods have been proposed for constructing confidence bands in regression models (see, for example, Brand *et al.* (1973) and Hauck

(1983) for the standard logistic regression model). Extending and evaluating these methods in the case of logistic regression with a cure fraction constitutes a stimulating topic for future research.

Acknowledgements

This research was supported by AIRES-Sud (AIRES-Sud is a program from the French Ministry of Foreign and European Affairs, implemented by the "Institut de Recherche pour le Développement", IRD-DSF), by the "Service de Coopération et d'Action Culturelle" of the French Embassy in Senegal, and by Edulink (program 9-ACP-RPR-118♯18). The authors also acknowledge grants from the "Ministère de la Recherche Scientifique" of Senegal.

References

Brand, R., Pinnock, D., and Jackson, K. (1973). Large sample confidence bands for the logistic response curve and its inverse. American Statistician **27**, 157–160.

Dietz, E. and Böhning, D., 2000. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics & Data Analysis **34**, 441-459.

Diop, A., Diop, A., and Dupuy, J.-F. (2011). Maximum likelihood estimation in the logistic regression model with a cure fraction. Electronic Journal of Statistics **5**, 460–483.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics **13**, 342–368.

Famoye, F. and Singh, K.P., 2006. Zero-inflated generalized Poisson regression model with an application to domestic violence data. Journal of Data Science **4**, 117-130.

Gouriéroux, C. and Monfort, A., 1981. Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. Journal of Econometrics **17**, 83–97.

Hall, D.B., 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics **56**, 1030–1039.

Hauck, W.W. (1983). A note on confidence bands for the logistic response curve. American Statistician **37**, 158–160.

Hilbe, J.M., 2009. Logistic regression models. Chapman & Hall: Boca Raton.

Hosmer, D.W. and Lemeshow, S., 2000. Applied logistic regression. Wiley: New York.

Kelley, M. E. and Anderson, S.J., 2008. Zero inflation in ordinal data: incorporating susceptibility to response through the use of a mixture model. Statistics in Medicine **27**, 3674–3688.

Lam, K.F., Xue, H. and Cheung, Y.B., 2006. Semiparametric analysis of zero-inflated count data. Biometrics **62**, 996–1003.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics **34**, 1–14.

Lee, A.H., Wang, K., Scott, J.A., Yau, K.K.W. and McLachlan, G. J., 2006. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. Statistical Methods in Medical Research **15**, 47-61.

Moghimbeigi, A., Eshraghian, M.R., Mohammad, K., and McArdle, B., 2009. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. Journal of Applied Statistics **35**, 1193–1202.

Moghimbeigi, A., Eshraghian, M.R., Mohammad, K., and McArdle, B., 2009. A score test for zero-inflation in multilevel count data. Computational Statistics & Data Analysis **53**, 1239–1248.

Ridout, M., Hinde, J., and Demétrio, C.G.B., 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics **57**, 219–223.

Xiang, L., Lee, A. H., Yau, K. K. W., and McLachlan, G. J., 2007. A score test for overdispersion in zero-inflated Poisson mixed regression model. Statistics in Medicine **26**, 1608–1622.

Table 1. Simulation results for model $\mathcal{M}_1$: $\beta = (-1.7, -2, -3.4, 5, .3)$ (percentage of infected among the susceptibles: 30%).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Percentage of immunes = 25%, $\theta = (.71, 1, 2, -3)$									
100	-1.688	-2.019	-3.417	4.648	0.322	0.667	1.154	2.153	-3.088
	(1.625)	(1.160)	(1.704)	(1.866)	(1.880)	(2.660)	(1.462)	(1.715)	(2.663)
	[1.330]	[0.689]	[1.387]	[1.525]	[1.524]	[2.199]	[1.161]	[1.410]	[2.156]
		0.631*	0.539*	0.316*	0.013*				
500	-1.710	-2.006	-3.392	4.991	0.311	0.677	1.076	2.079	-2.994
	(0.966)	(0.427)	(0.923)	(1.167)	(1.581)	(0.849)	(0.804)	(1.144)	(2.090)
	[0.734]	[0.330]	[0.686]	[0.907]	[1.329]	[0.711]	[0.643]	[0.910]	[1.663]
		0.997*	0.993*	0.984*	0.090*				
1000	-1.702	-2.004	-3.398	4.968	0.305	0.697	1.046	2.026	-2.997
	(0.579)	(0.297)	(0.584)	(0.797)	(0.843)	(0.749)	(0.623)	(0.779)	(1.127)
	[0.456]	[0.233]	[0.412]	[0.612]	[0.716]	[0.611]	[0.477]	[0.605]	[0.910]
		1*	0.998*	1*	0.107*				
1500	-1.720	-1.998	-3.410	4.971	0.305	0.709	1.035	2.013	-3.002
	(0.492)	(0.272)	(0.474)	(0.649)	(0.794)	(0.607)	(0.475)	(0.614)	(0.979)
	[0.384]	[0.206]	[0.332]	[0.484]	[0.675]	[0.487]	[0.361]	[0.484]	[0.778]
		1*	1*	1*	0.095*				
Percentage of immunes = 50%, $\theta = (-.3, -1, 2.1, 1)$									
100	-1.767	-2.105	-3.341	5.334	0.317	-0.377	-1.123	2.212	1.090
	(2.068)	(1.013)	(1.783)	(2.557)	(2.758)	(2.672)	(1.965)	(2.156)	(2.851)
	[1.682]	[0.784]	[1.464]	[2.150]	[2.312]	[2.141]	[1.265]	[1.760]	[2.366]
		0.335*	0.337*	0.127*	0*				
500	-1.716	-2.081	-3.576	5.217	0.294	-0.342	-1.092	2.097	1.078
	(1.417)	(0.545)	(0.930)	(1.704)	(1.704)	(1.079)	(0.760)	(1.518)	(1.746)
	[1.085]	[0.452]	[0.743]	[1.383]	[1.320]	[0.875]	[0.554]	[1.177]	[1.401]
		1*	1*	0.839*	0.057*				
1000	-1.701	-2.076	-3.529	5.015	0.304	-0.318	-1.056	2.105	1.031
	(0.780)	(0.391)	(0.627)	(1.072)	(1.113)	(0.743)	(0.473)	(1.028)	(1.160)
	[0.650]	[0.316]	[0.495]	[0.866]	[0.892]	[0.607]	[0.352]	[0.770]	[0.905]
		1*	1*	0.984*	0.045*				
1500	-1.698	-2.013	-3.482	5.007	0.303	-0.315	-1.023	2.103	1.024
	(0.694)	(0.294)	(0.489)	(0.857)	(0.865)	(0.609)	(0.384)	(0.885)	(0.926)
	[0.568]	[0.242	[0.381]	[0.694]	[0.685]	[0.497]	[0.290]	[0.642]	[0.721]
		1*	1*	0.999*	0.057*				
Percentage of immunes = 75%, $\theta = (.4, -1, -.6, -2)$									
100	-1.661	-2.131	-3.387	4.830	0.332	0.410	-1.059	-0.610	-2.158
	(2.139)	(2.127)	(2.803)	(3.283)	(3.661)	(3.554)	(1.995)	(3.118)	(2.974)
	[1.754]	[1.720]	[2.394]	[2.849]	[2.811]	[2.823]	[1.380]	[2.587]	[2.511]
		0.043*	0.064*	0.005*	0*				
500	-1.673	-2.075	-3.435	4.987	0.325	0.407	-1.060	-0.607	-1.921
	(1.436)	(1.012)	(1.614)	(2.455)	(2.198)	(1.295)	(0.598)	(1.651)	(1.884)
	[1.103]	[0.848]	[1.337]	[2.039]	[1.765]	[1.046]	[0.409]	[1.290]	[1.471]
		0.747*	0.787*	0.641*	0.046*				
1000	-1.545	-2.053	-3.399	5.024	0.309	0.405	-1.045	-0.604	-1.980
	(0.847)	(0.783)	(1.157)	(2.069)	(1.394)	(0.940)	(0.344)	(1.006)	(1.044)
	[0.669]	[0.619]	[0.899]	[1.586]	[1.125]	[0.737]	[0.253]	[0.775]	[0.843]
		0.994*	0.970*	0.895*	0.083*				
1500	-1.595	-2.017	-3.410	5.024	0.306	0.404	-1.035	-0.605	-1.997
	(0.708)	(0.630)	(0.909)	(0.895)	(1.234)	(0.749)	(0.259)	(0.866)	(0.860)
	[0.543]	[0.492]	[0.679]	[0.738]	[0.991]	[0.606]	[0.196]	[0.661]	[0.679]
		1*	0.993*	0.937*	0.089*				

Table 2. Simulation results for model $\mathcal{M}_2$: $\beta = (1.5, -2.3, 2.5, -3.5, .5)$ (percentage of infected among the susceptibles: 70%).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Percentage of immunes $= 25\%,\ \theta = (.71, 1, 2, -3)$									
100	1.512	-2.369	2.522	-3.518	0.562	0.725	1.191	2.061	-2.896
	(1.413)	(1.165)	(1.189)	(2.180)	(1.853)	(0.979)	(0.830)	(2.728)	(1.646)
	[1.190]	[0.917]	[0.949]	[1.810]	[1.582]	[0.789]	[0.636]	[2.309]	[1.264]
		0.800*	0.814*	0.271*	0.236*				
500	1.508	-2.313	2.520	-3.497	0.514	0.714	1.076	2.045	-2.959
	(0.905)	(0.635)	(0.568)	(1.198)	(0.633)	(0.419)	(0.434)	(1.617)	(0.652)
	[0.724]	[0.478]	[0.435]	[0.937]	[0.545]	[0.335]	[0.268]	[1.290]	[0.454]
		0.993*	0.993*	0.991*	0.629*				
1000	1.499	-2.297	2.508	-3.502	0.512	0.712	1.071	2.025	-2.985
	(0.569)	(0.488)	(0.398)	(0.908)	(0.557)	(0.308)	(0.387)	(1.178)	(0.365)
	[0.453]	[0.335]	[0.286]	[0.663]	[0.479]	[0.241]	[0.204]	[0.941]	[0.273]
		0.999*	0.998*	0.997*	0.732*				
1500	1.499	-2.299	2.497	-3.503	0.504	0.708	1.050	2.012	-2.985
	(0.339)	(0.372)	(0.322)	(0.701)	(0.522)	(0.257)	(0.337)	(0.983)	(0.289)
	[0.331]	[0.252]	[0.224]	[0.508]	[0.447]	[0.204]	[0.174]	[0.766]	[0.225]
		1*	1*	1*	0.764*				
Percentage of immunes $= 50\%,\ \theta = (-.3, -1, 2.1, 1)$									
100	1.526	-2.328	2.339	-3.336	0.488	-0.332	-1.107	2.179	1.053
	(1.887)	(1.824)	(2.170)	(2.570)	(2.181)	(0.902)	(1.204)	(1.946)	(1.567)
	[1.577]	[1.507]	[1.806]	[2.174]	[1.654]	[0.679]	[0.809]	[1.391]	[1.158]
		0.411*	0.228*	0.101*	0.060*				
500	1.517	-2.295	2.635	-3.397	0.537	-0.284	-0.983	2.127	1.041
	(0.956)	(0.772)	(0.687)	(1.352)	(0.826)	(0.317)	(0.472)	(0.466)	(1.043)
	[0.775]	[0.580]	[0.530]	[1.045]	[0.648]	[0.255]	[0.244]	[0.361]	[0.678]
		0.999*	0.963*	0.924*	0.339*				
1000	1.517	-2.303	2.563	-3.408	0.512	-0.296	-1.023	2.110	1.026
	(0.650)	(0.531)	(0.467)	(0.962)	(0.573)	(0.207)	(0.157)	(0.310)	(0.532)
	[0.518]	[0.390]	[0.364]	[0.701]	[0.454]	[0.169]	[0.123]	[0.243]	[0.366]
		1*	0.998*	0.999*	0.399*				
1500	1.498	-2.299	2.531	-3.365	0.513	-0.295	-1.012	2.112	0.995
	(0.473)	(0.389)	(0.355)	(0.768)	(0.455)	(0.181)	(0.122)	(0.260)	(0.339)
	[0.384]	[0.281]	[0.281]	[0.553]	[0.373]	[0.146]	[0.095]	[0.208]	[0.253]
		1*	1*	1*	0.451*				
Percentage of immunes $= 75\%,\ \theta = (.4, -1, -.6, -2)$									
100	1.489	-2.384	2.521	-3.458	0.554	0.420	-1.116	-0.557	-2.123
	(2.356)	(2.356)	(2.263)	(2.793)	(2.326)	(1.163)	(1.602)	(1.665)	(1.904)
	[1.934]	[1.979]	[1.944]	[2.389]	[1.902]	[0.875]	[1.121]	[1.232]	[1.450]
		0.109*	0.070*	0.106*	0.055*				
500	1.515	-2.381	2.523	-3.482	0.548	0.418	-1.047	-0.633	-2.084
	(1.282)	(1.405)	(1.701)	(1.910)	(1.757)	(0.394)	(0.396)	(0.651)	(0.665)
	[1.047]	[1.133]	[1.317]	[1.527]	[1.420]	[0.322]	[0.268]	[0.482]	[0.491]
		0.929*	0.899*	0.756*	0.353*				
1000	1.490	-2.295	2.522	-3.515	0.521	0.415	-1.031	-0.627	-1.997
	(0.862)	(0.916)	(1.141)	(1.228)	(1.282)	(0.243)	(0.185)	(0.345)	(0.398)
	[0.694]	[0.709]	[0.793]	[0.958]	[1.013]	[0.198]	[0.142]	[0.271]	[0.308]
		0.999*	0.999*	0.929*	0.367*				
1500	1.508	-2.304	2.497	-3.512	0.513	0.415	-1.023	-0.613	-1.997
	(0.721)	(0.683)	(0.867)	(0.957)	(0.961)	(0.221)	(0.149)	(0.293)	(0.299)
	[0.573]	[0.505]	[0.575]	[0.728]	[0.765]	[0.181]	[0.119]	[0.218]	[0.236]
		1*	1*	0.973*	0.359*				

Table 3. Simulation results for model $\mathcal{M}_3$: $\beta = (-1.7, -2.8, 0, -.7, 1.1)$ (percentage of infected among the susceptibles: 30%).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Percentage of immunes $= 25\%$, $\theta = (.71, 1, 2, -3)$									
100	-1.732	-2.918	0.017	-0.764	1.190	0.754	1.169	1.857	-2.975
	(1.630)	(0.631)	(1.310)	(1.721)	(1.857)	(2.681)	(1.686)	(1.983)	(2.877)
	[1.258]	[0.502]	[0.851]	[1.297]	[1.569]	[2.080]	[1.240]	[1.675]	[2.325]
		0.818*	0.033†	0.053*	0.052*				
500	-1.688	-2.908	-0.016	-0.751	1.114	0.729	1.164	2.139	-3.106
	(0.570)	(0.386)	(0.352)	(0.741)	(1.231)	(0.837)	(0.634)	(1.494)	(1.162)
	[0.441]	[0.312]	[0.196]	[0.465]	[0.956]	[0.669]	[0.449]	[1.224]	[0.898]
		0.998*	0.055†	0.301*	0.174*				
1000	-1.698	-2.853	-0.004	-0.713	1.079	0.726	1.080	2.083	-3.088
	(0.385)	(0.272)	(0.145)	(0.383)	(0.781)	(0.553)	(0.363)	(1.103)	(0.771)
	[0.304]	[0.217]	[0.115]	[0.301]	[0.608]	[0.442]	[0.269]	[0.913]	[0.596]
		1*	0.055†	0.494*	0.294*				
1500	-1.704	-2.837	-0.004	-0.705	1.110	0.716	1.057	2.071	-3.087
	(0.301)	(0.210)	(0.126)	(0.277)	(0.650)	(0.452)	(0.292)	(0.906)	(0.680)
	[0.234]	[0.168]	[0.101]	[0.224]	[0.489]	[0.359]	[0.205]	[0.743]	[0.503]
		1*	0.049†	0.707*	0.479*				
Percentage of immunes $= 50\%$, $\theta = (-.3, -1, 2.1, 1)$									
100	-1.776	-2.912	-0.039	-0.776	1.203	-0.336	-1.116	1.994	1.108
	(1.879)	(0.986)	(1.824)	(1.782)	(2.056)	(2.904)	(2.260)	(2.630)	(2.878)
	[1.542]	[0.807]	[1.404]	[1.493]	[1.709]	[2.143]	[1.533]	[2.053]	[2.129]
		0.472*	0.076†	0.008*	0.018*				
500	-1.753	-2.918	-0.030	-0.768	1.194	-0.279	-0.974	2.197	1.035
	(1.191)	(0.590)	(0.490)	(1.361)	(1.307)	(0.752)	(0.806)	(1.312)	(1.053)
	[0.919]	[0.481]	[0.371]	[1.056]	[1.021]	[0.590]	[0.456]	[0.929]	[0.747]
		1*	0.126†	0.108*	0.196*				
1000	-1.718	-2.853	0.005	-0.719	1.127	-0.288	-1.003	2.149	1.021
	(0.647)	(0.417)	(0.293)	(0.875)	(0.899)	(0.509)	(0.522)	(0.833)	(0.654)
	[0.525]	[0.335]	[0.224]	[0.682]	[0.710]	[0.414]	[0.259]	[0.591]	[0.495]
		1*	0.084†	0.148*	0.295*				
1500	-1.696	-2.824	-0.002	-0.705	1.117	-0.303	-1.020	2.119	1.021
	(0.551)	(0.329)	(0.310)	(0.669)	(0.701)	(0.387)	(0.304)	(0.561)	(0.490)
	[0.442]	[0.259]	[0.181]	[0.517]	[0.557]	[0.314]	[0.186]	[0.423]	[0.385]
		0.999*	0.078†	0.190*	0.383*				
Percentage of immunes $= 75\%$, $\theta = (.4, -1, -.6, -2)$									
100	-1.684	-2.948	-0.027	-0.792	1.215	0.497	-1.188	-0.587	-2.120
	(2.086)	(1.581)	(1.912)	(1.939)	(2.648)	(3.491)	(2.037)	(2.879)	(2.731)
	[1.689]	[1.313]	[1.591]	[1.621]	[2.224]	[2.297]	[1.493]	[2.183]	[2.215]
		0.127*	0.027†	0.013*	0.003*				
500	-1.774	-2.898	-0.028	-0.746	1.197	0.476	-0.923	-0.592	-1.905
	(1.392)	(0.908)	(0.993)	(1.651)	(1.898)	(0.952)	(1.042)	(1.396)	(1.664)
	[0.976]	[0.750]	[0.752]	[1.215]	[1.567]	[0.720]	[0.616]	[0.959]	[1.156]
		0.932*	0.162†	0.141*	0.084*				
1000	-1.745	-2.851	-0.004	-0.746	1.162	0.473	-0.925	-0.595	-1.927
	(0.731)	(0.631)	(0.587)	(0.934)	(1.430)	(0.648)	(0.797)	(0.742)	(0.978)
	[0.540]	[0.512]	[0.435]	[0.715]	[1.157]	[0.477]	[0.381]	[0.545]	[0.678]
		1*	0.125†	0.205*	0.143*				
1500	-1.733	-2.831	-0.004	-0.746	1.098	0.461	-0.970	-0.595	-1.960
	(0.596)	(0.514)	(0.360)	0.677	(1.289)	(0.513)	(0.447)	(0.572)	(0.845)
	[0.407]	[0.410]	[0.282]	[0.538]	[1.002]	[0.383]	[0.247]	[0.417]	[0.545]
		1*	0.114†	0.247*	0.186*				

Table 4. Simulation results for model $\mathcal{M}_4$: $\beta = (1.5, -2, 0, 3.5, -4)$ (percentage of infected among the susceptibles: 70%).

	$\widehat{\beta}_n$					$\widehat{\theta}_n$			
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$	$\widehat{\theta}_{1,n}$	$\widehat{\theta}_{2,n}$	$\widehat{\theta}_{3,n}$	$\widehat{\theta}_{4,n}$
Percentage of immunes $= 25\%$, $\theta = (.71, 1, 2, -3)$									
100	1.663	-2.215	-0.004	3.303	-3.512	0.793	0.939	2.216	-2.951
	(1.209)	(1.236)	(1.592)	(1.775)	(1.895)	(1.185)	(1.108)	(1.304)	(1.989)
	[1.005]	[0.975]	[1.125]	[1.377]	[1.567]	[0.926]	[0.856]	[1.050]	[1.604]
		0.364^*	$0.026^\dagger$	0.364^*	0.117^*				
500	1.580	-2.230	0.039	3.614	-3.776	0.671	1.132	2.130	-2.943
	(0.991)	(0.959)	(0.442)	(0.842)	(1.528)	(0.486)	(0.745)	(1.041)	(1.805)
	[0.793]	[0.660]	[0.339]	[0.679]	[1.108]	[0.391]	[0.454]	[0.713]	[1.324]
		0.954^*	$0.137^\dagger$	0.935^*	0.778^*				
1000	1.548	-2.096	0.028	3.559	-3.988	0.685	1.051	2.081	-2.954
	(0.664)	(0.712)	(0.256)	(0.565)	(1.098)	(0.326)	(0.419)	(0.654)	(1.241)
	[0.519]	[0.429]	[0.198]	[0.452]	[0.820]	[0.262]	[0.256]	[0.427]	[0.882]
		1^*	$0^\dagger$	1^*	0.909^*				
1500	1.514	-2.036	0.010	3.538	-3.997	0.707	1.031	2.031	-2.975
	(0.560)	(0.589)	(0.206)	(0.431)	(0.958)	(0.312)	(0.299)	(0.420)	(0.871)
	[0.429]	[0.314]	[0.161]	[0.349]	[0.697]	[0.246]	[0.192]	[0.318]	[0.669]
		0.965^*	$0.077^\dagger$	0.965^*	0.958^*				
Percentage of immunes $= 50\%$, $\theta = (-.3, -1, 2.1, 1)$									
100	1.472	-1.931	0.022	3.378	-3.427	-0.384	-0.913	2.286	1.276
	(1.790)	(1.663)	(2.130)	(1.998)	(2.231)	(0.979)	(1.487)	(1.607)	(1.563)
	[1.546]	[1.447]	[1.530]	[1.691]	[1.946]	[0.690]	[0.924]	[1.193]	[1.284]
		0.114^*	$0.184^\dagger$	0.163^*	0.005^*				
500	1.484	-2.034	0.004	3.428	-3.446	-0.339	-0.934	2.155	1.054
	(1.236)	(1.247)	(0.800)	(1.180)	(1.826)	(0.303)	(0.411)	(1.225)	(1.196)
	[1.020]	[0.936]	[0.518]	[0.999]	[1.505]	[0.248]	[0.232]	[0.633]	[0.945]
		0.878^*	$0.235^\dagger$	0.734^*	0.608^*				
1000	1.490	-1.956	0.008	3.475	-3.482	-0.325	-0.976	2.148	1.052
	(0.809)	(0.877)	(0.378)	(0.960)	(1.565)	(0.206)	(0.237)	(0.653)	(0.847)
	[0.652]	[0.603]	[0.282]	[0.812]	[1.187]	[0.166]	[0.130]	[0.401]	[0.658]
		0.994^*	$0.206^\dagger$	0.959^*	0.845^*				
1500	1.490	-1.987	-0.001	3.492	-3.763	-0.308	-0.982	2.092	1.032
	(0.570)	(0.637)	(0.322)	(0.784)	(0.989)	(0.178)	(0.208)	(0.475)	(0.649)
	[0.458]	[0.421]	[0.238]	[0.662]	[0.769]	[0.145]	[0.106]	[0.299]	[0.513]
		1^*	$0.172^\dagger$	0.989^*	0.892^*				
Percentage of immunes $= 75\%$, $\theta = (.4, -1, -.6, -2)$									
100	1.462	-1.936	-0.012	3.380	-3.520	0.508	-1.137	-0.710	-2.238
	(1.937)	(2.112)	(2.375)	(2.509)	(2.581)	(1.050)	(1.710)	(1.880)	(1.705)
	[1.643]	[1.790]	[1.833]	[2.211]	[2.146]	[0.791]	[1.018]	[1.329]	[1.445]
		0.042^*	$0.143^\dagger$	0.007^*	0.012^*				
500	1.456	-1.939	-0.020	3.453	-3.466	0.485	-0.933	-0.674	-2.156
	(1.493)	(1.418)	(0.820)	(1.900)	(2.061)	(0.478)	(0.410)	(0.644)	(1.395)
	[1.202]	[1.162]	[0.633]	[1.585]	[1.626]	[0.372]	[0.268]	[0.450]	[1.156]
		0.449^*	$0.297^\dagger$	0.116^*	0.231^*				
1000	1.477	-1.947	0.014	3.480	-3.785	0.462	-0.951	-0.645	-2.051
	(1.059)	(1.084)	(0.531)	(1.521)	(1.601)	(0.339)	(0.300)	(0.409)	(1.196)
	[0.851]	[0.857]	[0.419]	[1.222]	[1.273]	[0.259]	[0.170]	[0.291]	[0.969]
		0.912^*	$0.258^\dagger$	0.489^*	0.450^*				
1500	1.482	-1.975	-0.011	3.492	-3.801	0.437	-0.961	-0.642	-2.021
	(0.741)	(0.731)	(0.368)	(1.250)	(1.091)	(0.236)	(0.214)	(0.345)	(1.050)
	[0.597]	[0.561]	[0.290]	[0.969]	[0.905]	[0.190]	[0.121]	[0.242]	[0.838]
		0.971^*	$0.268^\dagger$	0.692^*	0.558^*				

Table 5. Simulation results for the model $\mathcal{M}_1$: $\beta = (-1.7, -2, -3.4, 5, .3)$ if there were no immunes.

n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$
100	-1.887	-2.474	-4.126	5.747	0.349
	(1.197)	(1.098)	(1.456)	(1.851)	(1.825)
	[0.938]	[0.750]	[1.031]	[1.437]	[1.418]
		0.985*	0.996*	0.953*	0.033*
500	-1.749	-2.072	-3.537	5.186	0.317
	(0.472)	(0.282)	(0.442)	(0.727)	(0.586)
	[0.366]	[0.217]	[0.332]	[0.558]	[0.469]
		1*	1*	1*	0.067*
1000	-1.724	-2.027	-3.449	5.066	0.302
	(0.318)	(0.188)	(0.275)	(0.456)	(0.436)
	[0.253]	[0.149]	[0.216]	[0.362]	[0.348]
		1*	1*	1*	0.121*
1500	-1.715	-2.020	-3.437	5.053	0.298
	(0.253)	(0.152)	(0.218)	(0.372)	(0.340)
	[0.199]	[0.121]	[0.169]	[0.297]	[0.273]
		1*	1*	1*	0.145*

Table 6. Simulation results for the model $\mathcal{M}_3$: $\beta = (-1.7, -2.8, 0, -.7, 1.1)$ if there were no immunes.

n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$
100	-1.881	-2.937	0.002	-0.753	1.293
	(0.830)	(0.550)	(0.385)	(0.969)	(1.057)
	[0.659]	[0.453]	[0.297]	[0.746]	[0.803]
		1*	0.044†	0.137*	0.244*
500	-1.740	-2.875	-0.002	-0.718	1.118
	(0.348)	(0.299)	(0.144)	(0.367)	(0.397)
	[0.272]	[0.228]	[0.115]	[0.289]	[0.314]
		1*	0.047†	0.540*	0.826*
1000	-1.728	-2.823	-0.001	-0.697	1.116
	(0.237)	(0.190)	(0.095)	(0.243)	(0.267)
	[0.188]	[0.151]	[0.078]	[0.197]	[0.212]
		1*	0.054†	0.809*	0.989*
1500	-1.711	-2.823	-0.001	-0.702	1.104
	(0.195)	(0.159)	(0.081)	(0.202)	(0.229)
	[0.154]	[0.125]	[0.065]	[0.162]	[0.184]
		1*	0.047†	0.935*	0.998*

Table 7. Results of a "naive" analysis of model $\mathcal{M}_1$: $\beta = (-1.7, -2, -3.4, 5, .3)$.

	$\widehat{\beta}_n$				
n	$\widehat{\beta}_{1,n}$	$\widehat{\beta}_{2,n}$	$\widehat{\beta}_{3,n}$	$\widehat{\beta}_{4,n}$	$\widehat{\beta}_{5,n}$
Percentage of immunes = 25%					
100	-1.093	-0.376	0.115	2.358	-1.288
	(3.448)	(3.792)	(4.844)	(5.860)	(4.619)
	[2.082]	[2.048]	[3.721]	[5.131]	[2.526]
		0.032*	0.294*	0.171*	0.497*
500	-1.376	-0.085	0.265	1.815	-1.572
	(3.609)	(1.929)	(4.058)	(5.159)	(2.845)
	[2.179]	[1.917]	[3.751]	[4.657]	[2.227]
		0.053*	0.559*	0.446*	0.701*
1000	-1.290	-0.158	0.158	2.410	-0.966
	(2.944)	(2.182)	(3.565)	(4.871)	(2.760)
	[2.019]	[1.921]	[3.558]	[4.304]	[1.843]
		0.046*	0.624*	0.535*	0.751*
1500	-1.171	-0.112	0.162	1.962	-1.044
	(2.138)	(1.902)	(3.570)	(4.885)	(2.659)
	[1.867]	[1.890]	[3.562]	[4.437]	[1.847]
		0.052*	0.652*	0.570*	0.778*
Percentage of immunes = 50%					
100	-1.953	-0.228	-1.048	3.165	0.171
	(3.843)	(4.118)	(5.234)	(5.194)	(0.506)
	[1.745]	[2.181]	[3.101]	[4.347]	[0.397]
		0.028*	0.508*	0.437*	0.139*
500	-1.611	-0.237	-0.663	2.754	0.426
	(1.583)	(2.603)	(3.193)	(5.193)	(0.458)
	[1.231]	[1.902]	[2.845]	[4.383]	[0.392]
		0.046*	0.704*	0.658*	0.457*
1000	-1.755	-0.148	-0.617	1.274	0.432
	(1.348)	(1.871)	(2.836)	(4.088)	(0.434)
	[1.175]	[1.855]	[2.800]	[3.725]	[0.371]
		0.043*	0.759*	0.731*	0.587*
1500	-1.446	-0.124	-0.621	1.425	0.487
	(1.281)	(1.891)	(2.800)	(3.632)	(0.433)
	[1.143]	[1.878]	[2.778]	[3.574]	[0.378]
		0.051*	0.799*	0.755*	0.632*
Percentage of immunes = 75%					
100	-1.665	-0.284	-1.095	1.434	0.178
	(4.897)	(4.546)	(5.127)	(6.948)	(6.963)
	[2.514]	[2.293]	[2.981]	[5.974]	[2.119]
		0.037*	0.515*	0.083*	0.237*
500	-1.746	-0.484	-1.032	0.751	0.446
	(3.847)	(3.748)	(4.757)	(6.987)	(6.769)
	[2.091]	[2.028]	[2.936]	[5.969]	[1.901]
		0.041*	0.696*	0.305*	0.596*
1000	-1.520	-0.261	-0.745	0.252	0.321
	(2.883)	(4.496)	(2.679)	(5.839)	(3.468)
	[1.857]	[2.110]	[2.655]	[5.662]	[1.194]
		0.038*	0.788*	0.481*	0.696*
1500	-1.510	-0.120	-0.757	0.132	0.315
	(2.550)	(1.902)	(2.671)	(6.016)	(3.658)
	[1.791]	[1.887]	[2.649]	[5.659]	[1.427]
		0.041*	0.801*	0.567*	0.727*

Table 8. Estimated probability $p(\mathbf{x})$ for the models $\mathcal{M}_1$ ($p(\mathbf{x}) = 0.250$) and $\mathcal{M}_2$ ($p(\mathbf{x}) = 0.343$).

	0% of immune		25% of immune		50% of immune		75% of immune	
n	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_1$	$\mathcal{M}_2$
100	0.229	0.335	0.267	0.363	0.264	0.353	0.266	0.362
	(0.118)	(0.117)	(0.191)	(0.164)	(0.264)	(0.215)	(0.360)	(0.367)
	[0.097]	[0.095]	[0.149]	[0.132]	[0.202]	[0.174]	[0.293]	[0.318]
	$0.964^{\star}$	$0.959^{\star}$	$0.964^{\star}$	$0.978^{\star}$	$0.987^{\star}$	$0.949^{\star}$	$0.951^{\star}$	$0.831^{\star}$
	$0.407^{\mp}$	$0.424^{\mp}$	$0.596^{\mp}$	$0.523^{\mp}$	$0.719^{\mp}$	$0.627^{\mp}$	$0.859^{\mp}$	$0.756^{\mp}$
500	0.246	0.343	0.261	0.356	0.263	0.358	0.255	0.354
	(0.046)	(0.049)	(0.079)	(0.069)	(0.110)	(0.084)	(0.188)	(0.201)
	[0.037]	[0.039]	[0.056]	[0.054]	[0.083]	[0.068]	[0.150]	[0.160]
	$0.957^{\star}$	$0.959^{\star}$	$0.930^{\star}$	$0.898^{\star}$	$0.921^{\star}$	$0.943^{\star}$	$0.891^{\star}$	$0.701^{\star}$
	$0.182^{\mp}$	$0.192^{\mp}$	$0.216^{\mp}$	$0.228^{\mp}$	$0.343^{\mp}$	$0.319^{\mp}$	$0.572^{\mp}$	$0.482^{\mp}$
1000	0.247	0.342	0.255	0.352	0.258	0.351	0.254	0.350
	(0.034)	(0.035)	(0.051)	(0.052)	(0.071)	(0.061)	(0.137)	(0.134)
	[0.027]	[0.028]	[0.037]	[0.039]	[0.055]	[0.049]	[0.106]	[0.106]
	$0.948^{\star}$	$0.954^{\star}$	$0.887^{\star}$	$0.875^{\star}$	$0.931^{\star}$	$0.932^{\star}$	$0.894^{\star}$	$0.651^{\star}$
	$0.128^{\mp}$	$0.136^{\mp}$	$0.149^{\mp}$	$0.164^{\mp}$	$0.248^{\mp}$	$0.226^{\mp}$	$0.436^{\mp}$	$0.359^{\mp}$
1500	0.249	0.343	0.252	0.348	0.253	0.352	0.251	0.348
	(0.028)	(0.028)	(0.037)	(0.038)	(0.062)	(0.047)	(0.106)	(0.108)
	[0.022]	[0.023]	[0.028]	[0.028]	[0.044]	[0.038]	[0.085]	[0.084]
	$0.945^{\star}$	$0.958^{\star}$	$0.907^{\star}$	$0.877^{\star}$	$0.932^{\star}$	$0.929^{\star}$	$0.904^{\star}$	$0.641^{\star}$
	$0.105^{\mp}$	$0.112^{\mp}$	$0.119^{\mp}$	$0.133^{\mp}$	$0.199^{\mp}$	$0.185^{\mp}$	$0.373^{\mp}$	$0.301^{\mp}$

Note: $(\cdot)$: root mean square error. $[\cdot]$: mean absolute error. *: empirical coverage probability. $^{\mp}$: average length of confidence intervals. For each percentage of immunes, the percentages of infected among the susceptibles are respectively 30% ($\mathcal{M}_1$) and 70% ($\mathcal{M}_2$). All results are based on 1500 replicates.

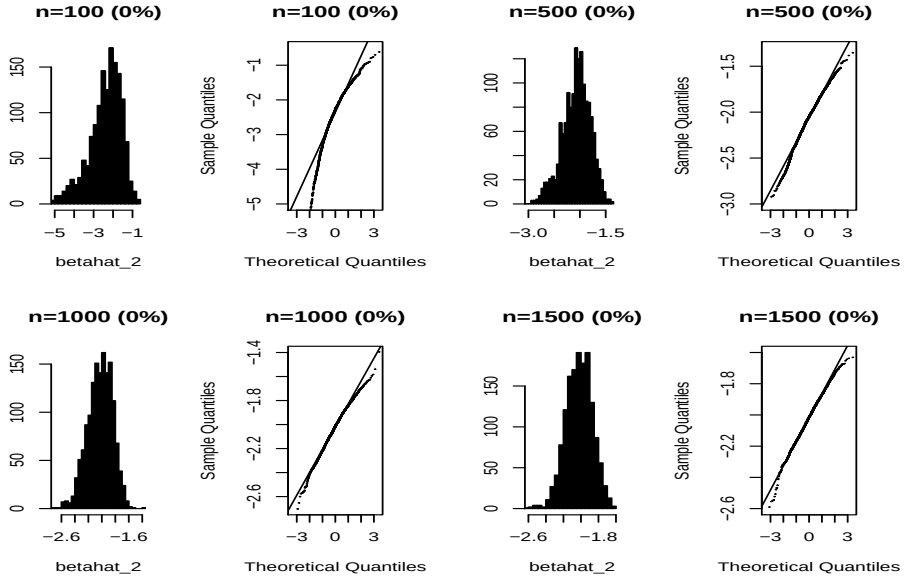

Figure 1. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with no immunes in the sample (the percentage of immunes is given in brackets). n is the sample size. All results are based on 1500 simulated datasets.

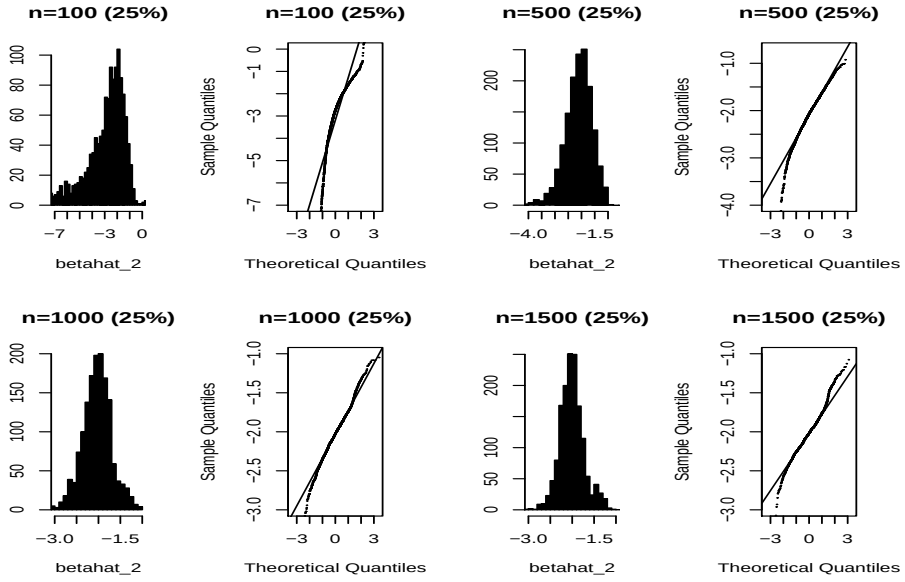

Figure 2. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with 25% of immunes.

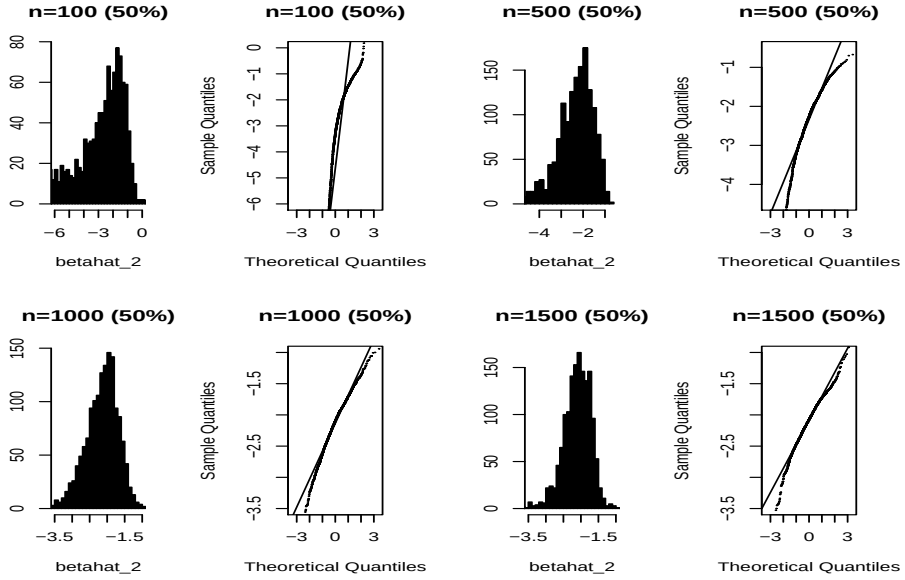

Figure 3. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with 50% of immunes.

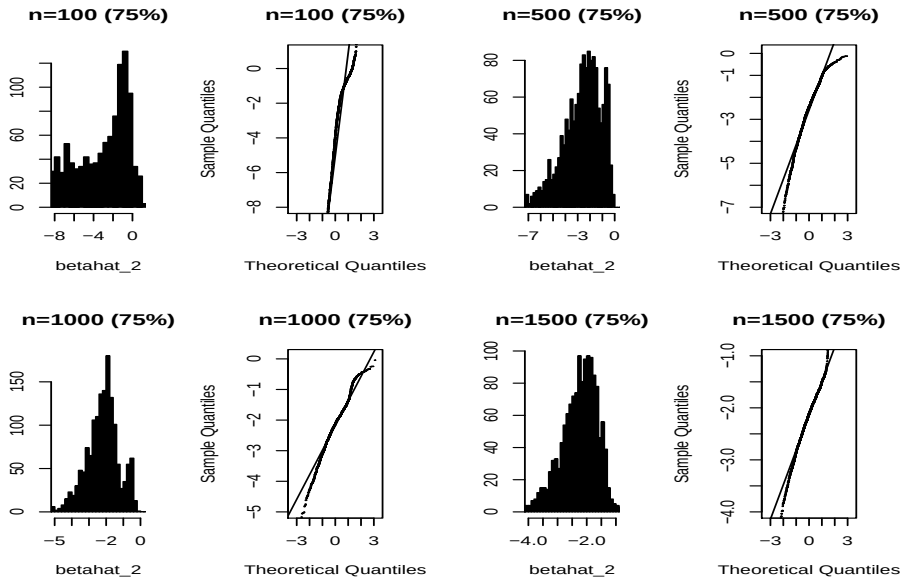

Figure 4. Histograms and Q-Q plots for $\widehat{\beta}_{2,n}$ in model $\mathcal{M}_1$, with 75% of immunes.

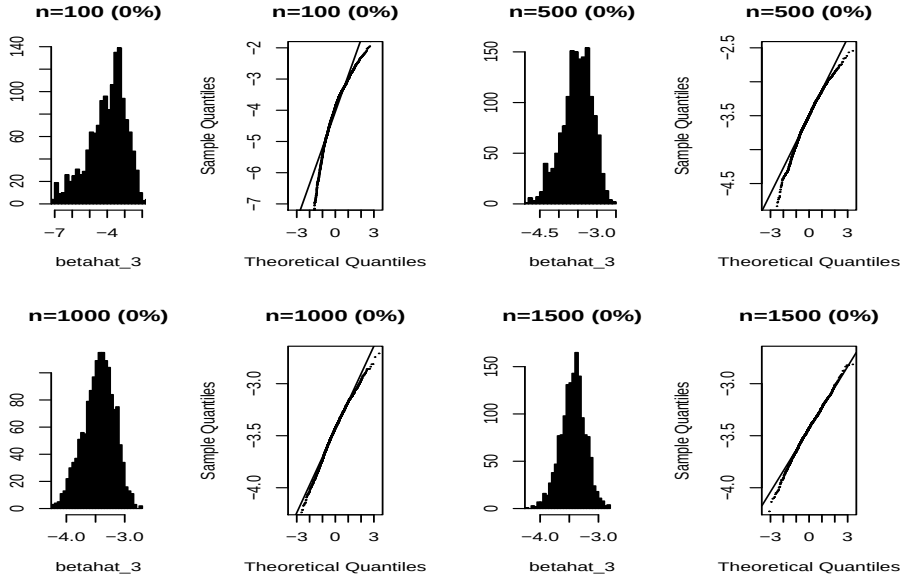

Figure 5. Histograms and Q-Q plots for $\widehat{\beta}_{3,n}$ in model $\mathcal{M}_1$, with no immunes in the sample.

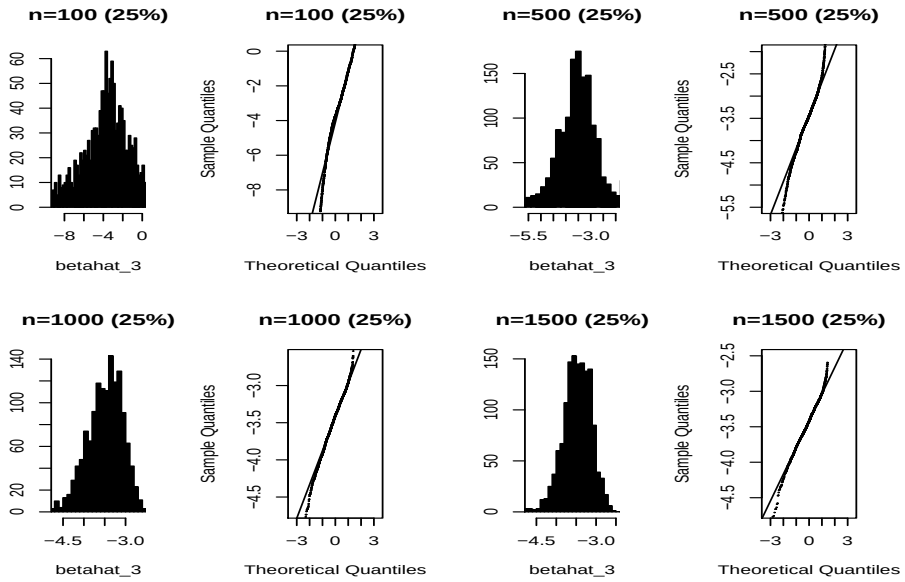

Figure 6. Histograms and Q-Q plots for $\widehat{\beta}_{3,n}$ in model $\mathcal{M}_1$, with 25% of immunes.

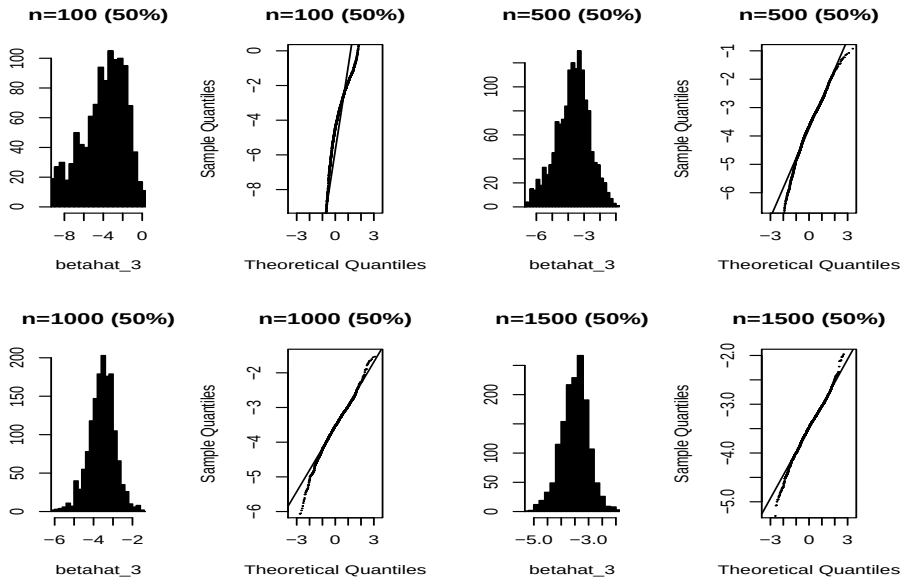

Figure 7. Histograms and Q-Q plots for $\widehat{\beta}_{3,n}$ in model $\mathcal{M}_1$, with 50% of immunes.

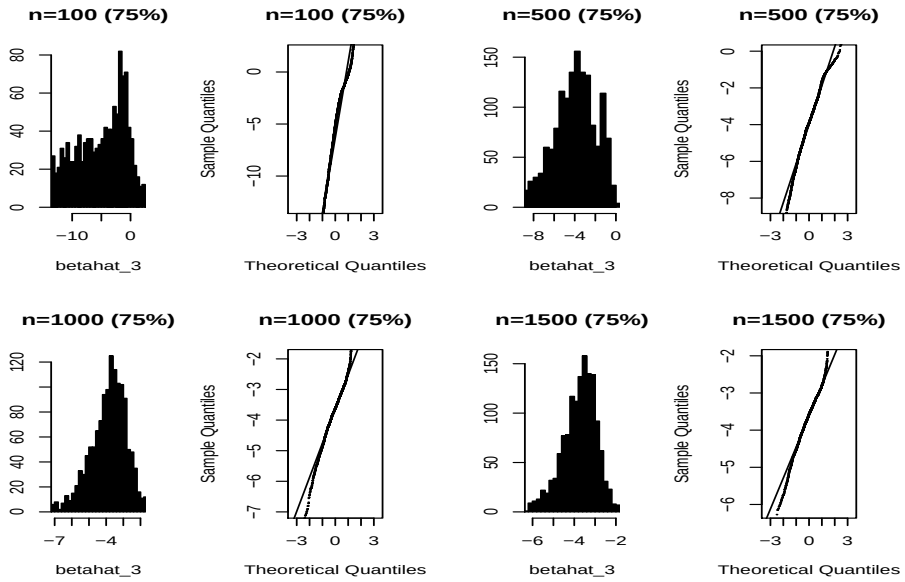

Figure 8. Histograms and Q-Q plots for $\widehat{\beta}_{3,n}$ in model $\mathcal{M}_1$, with 75% of immunes.

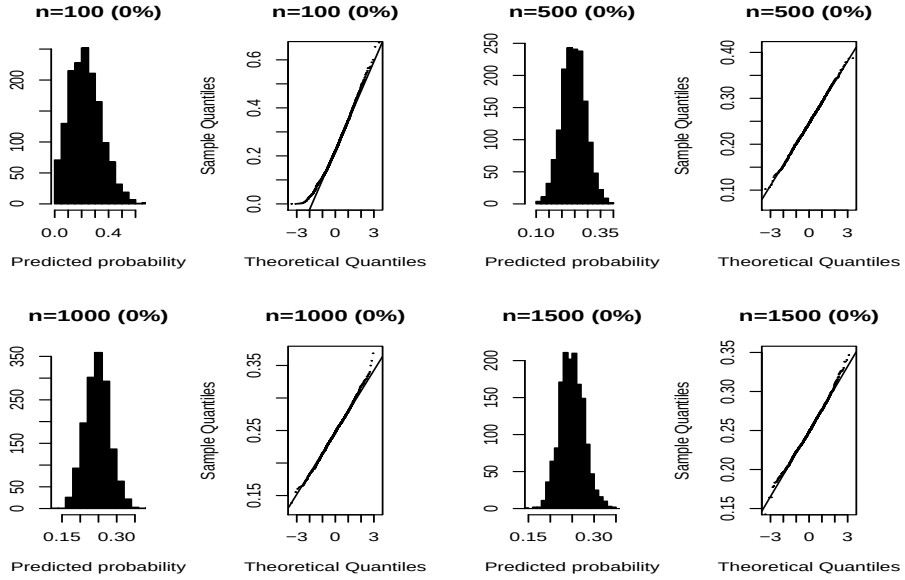

Figure 9. Histograms and Q-Q plots for $\widehat{p}_n(\mathbf{x})$ in model $\mathcal{M}_1$, with no immunes in the sample.

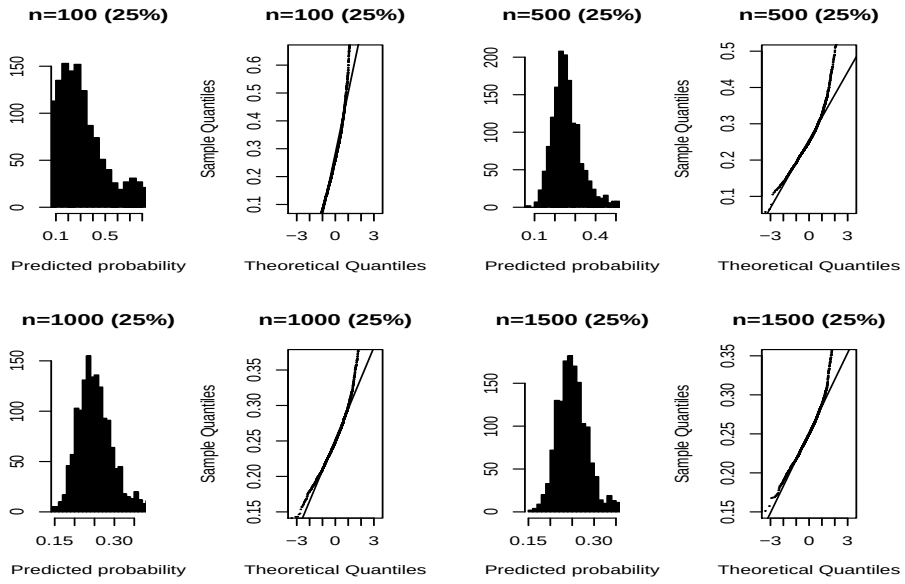

Figure 10. Histograms and Q-Q plots for $\widehat{p}_n(\mathbf{x})$ in model $\mathcal{M}_1$, with 25% of immunes.

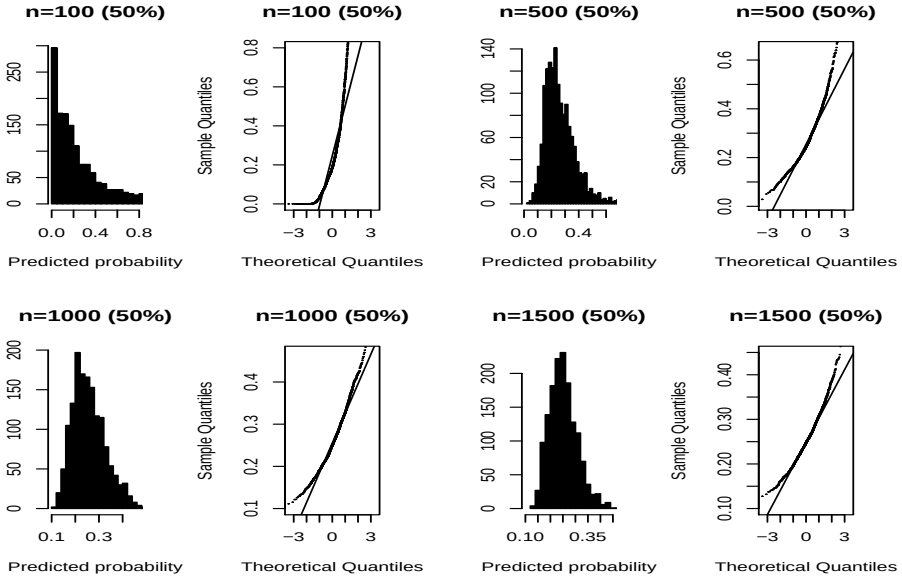

Figure 11. Histograms and Q-Q plots for $\widehat{p}_n(\mathbf{x})$ in model $\mathcal{M}_1$, with 50% of immunes.

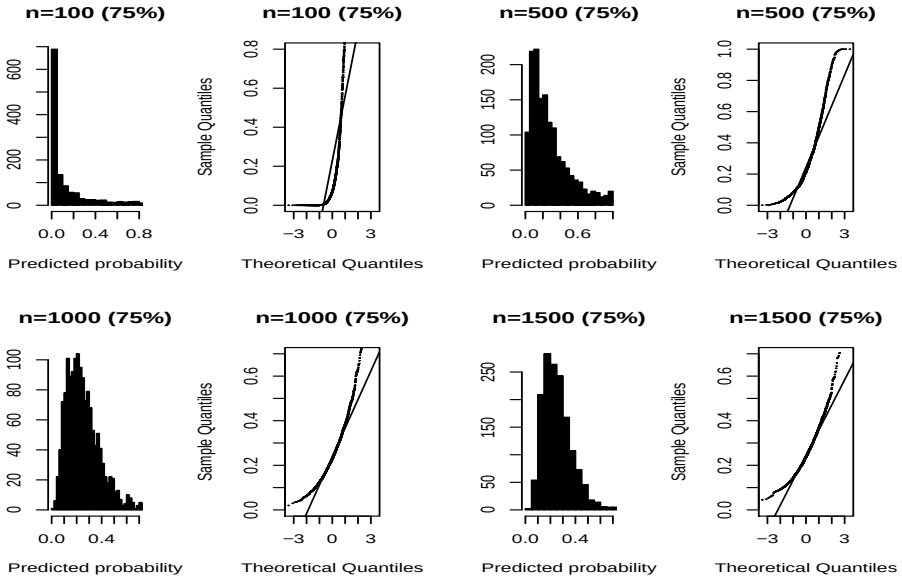

Figure 12. Histograms and Q-Q plots for $\widehat{p}_n(\mathbf{x})$ in model $\mathcal{M}_1$, with 75% of immunes.